Reversible Christoffel factorizations

Michelangelo Bucci[a], Alessandro De Luca[b,*], Luca Q. Zamboni[a,c]

[a] *Department of Mathematics, University of Turku, FI-20014 Turku, Finland*
[b] *DIETI, Università degli Studi di Napoli Federico II*
via Claudio 21, 80125 Napoli, Italy
[c] *Université de Lyon, Université Lyon 1, CNRS UMR 5208 Institut Camille Jordan*
Bâtiment du Doyen Jean Braconnier, 43, blvd du 11 novembre 1918
F-69622 Villeurbanne Cedex, France

Abstract

We define a family of natural decompositions of Sturmian words in Christoffel words, called *reversible Christoffel* (RC) factorizations. They arise from the observation that two Sturmian words with the same language have (almost always) arbitrarily long Abelian equivalent prefixes. Using the three gap theorem, we prove that in each RC factorization, only 2 or 3 distinct Christoffel words may occur. We begin the study of such factorizations, considered as infinite words over 2 or 3 letters, and show that in the general case they are either Sturmian words, or obtained by a three-interval exchange transformation.

Keywords: Sturmian word, Christoffel word, reversible Christoffel factorization, three-interval exchange transformation
2010 MSC: 68R15

1. Introduction

In combinatorics on words and symbolic dynamics, it is often meaningful to look at two infinite words $\mathbf{w}, \mathbf{w}'$ and determine the segments where they coincide, that is, locate maximal occurrences of factors u such that $\mathbf{w} = pu\mathbf{s}$, $\mathbf{w}' = p'u\mathbf{s}'$ for some words p and p' of equal length and some infinite words $\mathbf{s}, \mathbf{s}'$.

If $\mathbf{w}$ and $\mathbf{w}'$ are two fixed points of an irreducible Pisot substitution φ, the *strong coincidence conjecture* (proved by Barge and Diamond [1] in the binary case) states that there exists a letter a and two factorizations

$$\mathbf{w} = pa\mathbf{s}, \quad \mathbf{w}' = p'a\mathbf{s}' \tag{1}$$

such that p and p' are Abelian equivalent, i.e., an "anagram" of each other. This has two remarkable consequences:

1. $\mathbf{w}$ and $\mathbf{w}'$ agree on arbitrarily long segments (as defined above), i.e., they are *proximal*,

*Corresponding author
Email addresses: michelangelo.bucci@utu.fi (Michelangelo Bucci), alessandro.deluca@unina.it (Alessandro De Luca), lupastis@gmail.com (Luca Q. Zamboni)

Preprint submitted to Theoretical Computer Science

2. $\mathbf{w}$ and $\mathbf{w}'$ have arbitrarily long Abelian equivalent prefixes; in short, we say that they are *Abelian comparable*.

Indeed, from (1) it follows $\mathbf{w} = \varphi^n(p)\varphi^n(a)\varphi^n(\mathbf{s})$ and $\mathbf{w}' = \varphi^n(p')\varphi^n(a)\varphi^n(\mathbf{s}')$ for all $n \geq 0$.

This induces two factorizations (*comparison*)

$$\begin{aligned} \mathbf{w} &= x_1 x_2 \cdots x_n \cdots \\ \mathbf{w}' &= x'_1 x'_2 \cdots x'_n \cdots \end{aligned} \tag{2}$$

defined so that each pair of Abelian equivalent prefixes of $\mathbf{w}$ and $\mathbf{w}'$ is $(x_1 \cdots x_k, x'_1 \cdots x'_k)$ for some $k \geq 0$; equivalently, for all $k > 0$, x_k and x'_k are the shortest nonempty Abelian equivalent prefixes of the infinite words $x_k x_{k+1} \cdots$ and $x'_k x'_{k+1} \cdots$.

In this paper (a preliminary version of which was presented at the first RuFiDiM [2]), we look at Sturmian words over $A = \{0, 1\}$ from a similar point of view. We recall that an infinite word over A is *Sturmian* if it has exactly $n+1$ distinct factors of each length $n \geq 0$. The first systematic study of Sturmian sequences is usually credited to Morse and Hedlund [3], from the point of view of symbolic dynamics. In fact, every Sturmian word can be realized either as a *lower mechanical word* or as an *upper* one. The lower (resp. upper) mechanical word $\mathbf{s}_{\alpha,\rho}$ (resp. $\mathbf{s}'_{\alpha,\rho}$) of *slope* α and *intercept* ρ, with $0 \leq \alpha, \rho < 1$, is the infinite word indexed over $\mathbb{N}$ whose n-th letter is 0 if

$$\{n\alpha + \rho\} < 1 - \alpha \qquad \text{(resp. if } 0 < \{n\alpha + \rho\} \leq 1 - \alpha)$$

and 1 otherwise (denoting by $\{\sigma\}$ the *fractional part* $\sigma - \lfloor \sigma \rfloor$ of the real number σ). Thus mechanical words encode *rotations* by angle $2\pi\alpha$ on a circle, and α gives the frequency of the letter 1 in the infinite word. Moreover, the slope determines the *language* (set of factors). Among binary words, mechanical words are characterized by the *balance* property: the number of occurrences of the letter 1 in two factors of the same length may differ at most by 1.

When $\alpha \notin \mathbb{Q}$, the word $\mathbf{s}_{\alpha,\alpha} = \mathbf{s}'_{\alpha,\alpha} =: \mathbf{c}$ is said to be the *characteristic* (or *standard*) Sturmian word of slope α. Its prefixes are exactly all *left special* factors of the Sturmian words of slope α, i.e., $p \in \mathrm{Pref}(\mathbf{c})$ if and only if $0p, 1p \in \mathrm{Fact}(\mathbf{s}_{\alpha,\rho})$ for any ρ. We say that a Sturmian word is *singular* if it contains the characteristic word (of the same slope) as a proper suffix. By definition, we have

$$\mathbf{s}_{\alpha,0} = 0\mathbf{c} \quad \text{and} \quad \mathbf{s}'_{\alpha,0} = 1\mathbf{c}\ ;$$

any other singular Sturmian word can be written as $\tilde{p}01\mathbf{c}$ or $\tilde{p}10\mathbf{c}$, where $p \in \mathrm{Pref}(\mathbf{c})$. It is easy to see that every nonsingular Sturmian word is both an upper mechanical word and a lower one.

Mechanical words of irrational slope are exactly all Sturmian words, whereas those of rational slope are periodic words; when $\alpha \in \mathbb{Q}$, the shortest v such that $\mathbf{s}_{\alpha,0} = v^\omega$ (resp. $\mathbf{s}'_{\alpha,0} = v^\omega$) is the *lower* (resp. *upper*) *Christoffel word* of slope α. It is well-known (cf. [4]) that the set of lower (resp. upper) Christoffel words can be characterized as $A \cup 0\mathcal{P}1$ (resp. $A \cup 1\mathcal{P}0$), where $\mathcal{P}$ is the set of *central words*, i.e., words u such that both $0u1$ and $1u0$ are factor of some Sturmian word s. Central words are exactly all palindromic prefixes of characteristic Sturmian words. The following characterization of central words is well-known.

Proposition 1.1 (See de Luca [5]). *A word $u \in A^*$ is central if and only if it is a* palindrome *satisfying* one *of the following conditions:*

1. *u is a power of a letter, i.e., $u \in 0^* \cup 1^*$, or*

2. $u = p01q = q10p$ *for some palindromes* $p, q \in A^*$.

In the latter case, the words p, q *are central too, and uniquely determined; one of them is the longest palindromic prefix (and suffix) of* u*.*

By iterated application of the previous result, one easily obtains:

Corollary 1.2. *Let* v *be a proper palindromic prefix of a central word* u*, such that* $v \notin 0^* \cup 1^*$*. Then either* $v01$ *or* $v10$ *is a prefix of* u*.*

Since central words are palindromes, we have that $v = 0u1$ is a lower Christoffel word if and only if its *reversal* $\tilde{v} = 1u0$ is an upper Christoffel word. We recall that any nontrivial (i.e., longer than a letter) Christoffel word can be uniquely written as a product of two (shorter) Christoffel words. All pairs (u, v) of lower Christoffel words such that uv is Christoffel make up the *Christoffel tree* (cf. Berstel and de Luca [4]; see Figure 1) where the pair $(0, 1)$ is the root[1], and every node (u, v) has the two children (u, uv) and (uv, v). Moreover, all lower Christoffel factors of an infinite Sturmian word are found on an infinite path on the tree. Thus, in particular,

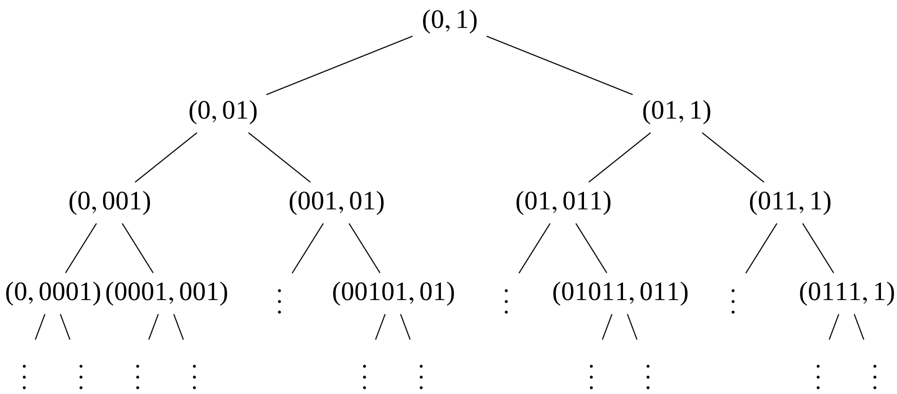

Figure 1: The Christoffel tree

the following properties hold:

Proposition 1.3. *Let* u, v, z *be Christoffel words such that* $z = uv$*, and let* $\mathbf{w}$ *be a Sturmian word with* $z \in \mathrm{Fact}(\mathbf{w})$*. Then:*

1. u^2v *and* uv^2 *are Christoffel words too, and exactly one of them is a factor of* $\mathbf{w}$*;*
2. *if* $\{u, v\} \neq \{0, 1\}$*, then either* u *is a prefix of* v*, or* v *is a suffix of* u*.*

For more information on Sturmian and Christoffel words, we refer the reader to [6, 7].

In the next sections, we shall deal with the comparison of Sturmian words. We begin by proving (Proposition 2.1) that all pairs of Sturmian words of the same slope, except one, are Abelian comparable. Our main result (Theorem 3.2) shows that for Abelian comparable Sturmian words $\mathbf{w}$ and $\mathbf{w}'$, the factorizations in (2) have at most 3 distinct terms, i.e., the sets $\{x_i\}_{i>0}$ and $\{x'_i\}_{i>0}$ have cardinality 2 or 3; furthermore, all such terms are Christoffel words. Finally, we shall examine the structure of these factorizations.

[1] Using $(1, 0)$ instead, all *upper* Christoffel pairs are obtained.

2. Comparison of Sturmian words: RC factorizations

Trivially, if two Sturmian words are Abelian comparable then they have the same slope (and hence the same language). The following proposition shows that the converse holds too, with a single exception.

Proposition 2.1. *Let* $\mathbf{w}, \mathbf{w}'$ *be Sturmian words of slope* α*, and* $\mathbf{c} = \mathbf{s}_{\alpha,\alpha}$ *be the characteristic word. If* $\{\mathbf{w}, \mathbf{w}'\} \neq \{0\mathbf{c}, 1\mathbf{c}\}$*, then* $\mathbf{w}$ *and* $\mathbf{w}'$ *are Abelian comparable.*

Proof. Let x_1 and x_1' be the shortest nonempty prefixes (of $\mathbf{w}$ and $\mathbf{w}'$ respectively) which are Abelian equivalent. These are well defined; in fact, suppose by contradiction that $\mathbf{w}$ and $\mathbf{w}'$ have no Abelian equivalent prefixes except ε. By the balance property, it follows $\{\mathbf{w}, \mathbf{w}'\} = \{0\mathbf{t}, 1\mathbf{t}\}$ for some infinite word $\mathbf{t}$; as all prefixes of $\mathbf{t}$ are left special, we get $\mathbf{t} = \mathbf{c}$, contradicting our hypothesis.

Let then $\mathbf{w} = x_1\mathbf{w}_1$ and $\mathbf{w}' = x_1'\mathbf{w}_1'$ for some Sturmian words $\mathbf{w}_1, \mathbf{w}_1'$ having the same language as $\mathbf{w}$ and $\mathbf{w}'$. We have $\{\mathbf{w}_1, \mathbf{w}_1'\} \neq \{0\mathbf{c}, 1\mathbf{c}\}$, for otherwise we would obtain $\{\mathbf{w}, \mathbf{w}'\} = \{\tilde{p}10\mathbf{c}, \tilde{p}01\mathbf{c}\}$ for some $p \in \operatorname{Pref}(\mathbf{c})$, and then $\{x_1, x_1'\} = \{\tilde{p}1, \tilde{p}0\}$, which is absurd as x_1 and x_1' are Abelian equivalent. Hence we can iterate this argument to get infinitely many Abelian equivalent prefixes of $\mathbf{w}$ and $\mathbf{w}'$. ☐

Example 2.2. Let $\alpha = (3 - \sqrt{5})/2$. The *Fibonacci word*

$$\mathbf{f} = \mathbf{s}_{\alpha,\alpha} = 0100101001001010010100100101001001010010100100101001001\cdots$$

is the most famous Sturmian word. The Abelian comparison of the words $\mathbf{f}$ and $\mathbf{f}' := \mathbf{s}_{\alpha,4/5}$ is:

$$\begin{array}{ll|l|l|l|l|l|l|l|l|l|l|l|l|l|l|l|l|l|l|l|l}
\mathbf{f} & = 01 & 0 & 01 & 01 & 0 & 01 & 001 & 01 & 0 & 01 & 01 & 0 & 01 & 0 & 01 & 01 & 0 & 01 & 001 & 01 & \cdots \\
\mathbf{f}' & = 10 & 0 & 10 & 10 & 0 & 10 & 100 & 10 & 0 & 10 & 10 & 0 & 10 & 0 & 10 & 10 & 0 & 10 & 100 & 10 & \cdots
\end{array}$$

Let $\mathbf{w}, \mathbf{w}'$ be two Sturmian words having the same language, and suppose $\{\mathbf{w}, \mathbf{w}'\} \neq \{0\mathbf{c}, 1\mathbf{c}\}$ where $\mathbf{c}$ is the characteristic word with the same slope. By Proposition 2.1, $\mathbf{w}$ and $\mathbf{w}'$ are then Abelian comparable; let their comparison be given by (2).

By definition, x_i and x_i' are Abelian equivalent for all $i \geq 1$. By the balance property, it follows either $x_i = x_i' \in A$, or $\{x_i, x_i'\} = \{0u1, 1u0\}$ for some factor u of $\mathbf{w}$, which is then a central word. We conclude that in all cases, x_i and x_i' are Christoffel words, with $x_i' = \widetilde{x_i}$. Thus we can write:

$$\begin{aligned} \mathbf{w} &= x_1 x_2 \cdots x_n \cdots , \\ \mathbf{w}' &= \widetilde{x_1}\widetilde{x_2} \cdots \widetilde{x_n} \cdots . \end{aligned} \tag{3}$$

Conversely, if $(x_n)_{n>0}$ is a sequence of Christoffel words such that both infinite words in (3) are Sturmian, then the Abelian comparison of $\mathbf{w}$ and $\mathbf{w}'$ yields exactly the same factorizations.

This motivates the following definition: we call *reversible Christoffel (RC) factorization* of a Sturmian word $\mathbf{w}$ any infinite sequence $(x_k)_{k>0}$ of Christoffel words such that

1. $\mathbf{w} = x_1 x_2 \cdots x_n \cdots$, and
2. $\mathbf{w}' := \widetilde{x_1}\widetilde{x_2} \cdots \widetilde{x_n} \cdots$ is a Sturmian word.

We also say that $(x_k)_{k>0}$ is *the* RC factorization of $\mathbf{w}$ *determined by* $\mathbf{w}'$.

A trivial RC factorization is obtained by choosing all x_k's to be single letters, so that $\mathbf{w}' = \mathbf{w}$. The definition implies that every choice of $\mathbf{w}'$ determines a distinct factorization of $\mathbf{w}$; this proves the following statement.

Proposition 2.3. *Every Sturmian word admits uncountably many distinct RC factorizations.*

The following result is an immediate consequence of the balance property:

Proposition 2.4. *Let* $\mathbf{w}$ *be a Sturmian word, and* $(x_k)_{k>0}$ *be any RC factorization of* $\mathbf{w}$*. Then the terms* x_k*,* $k \geq 1$*, are either* all *upper Christoffel words, or they are all lower Christoffel words.*

In the remainder of this section, we shall use the above definitions and results to give a proof of a stronger (and probably known) version of the strong coincidence conjecture in the case of Sturmian words, namely Proposition 2.7 below. We need the following lemma (a restatement of [8, Lemma 5.2]):

Lemma 2.5 (See Bucci et al. [8])**.** *Two distinct Sturmian words* $\mathbf{w}, \mathbf{w}'$ *are proximal if and only if they can be written as* $\mathbf{w} = q\mathbf{c}$*,* $\mathbf{w}' = q'\mathbf{c}$ *for some* q, q' *with* $|q| = |q'|$*; that is, if and only if they contain the characteristic word at the same position.*

As observed above, in such a case the set $\{q, q'\}$ is either $\{0, 1\}$ or $\{\tilde{p}01, \tilde{p}10\}$ for some $p \in \mathrm{Pref}(\mathbf{c})$.

We recall that a morphism $f : A^* \to A^*$ is said to be *Sturmian* if it maps Sturmian words to Sturmian words; as is well-known (cf. [6]), for a morphism to be Sturmian it suffices to map *one* Sturmian word to another one. Furthermore, Sturmian morphisms form a monoid generated by the three substitutions

$$E : 0 \mapsto 1, 1 \mapsto 0, \quad \varphi : 0 \mapsto 01, 1 \mapsto 0, \quad \text{and } \tilde{\varphi} : 0 \mapsto 10, 1 \mapsto 0\ . \tag{4}$$

From this, an easy induction argument gives the following property, also well-known:

Lemma 2.6. *If* $f \notin \{\mathrm{id}, E\}$ *is a Sturmian morphism, then one between* $f(0)$ *and* $f(1)$ *is a proper prefix or a proper suffix of the other.*

We also recall that a complete characterization of Sturmian fixed points of morphisms in terms of slope and intercept was given by Yasutomi [9] (see also [10]).

The following result about coincidence for Sturmian words can be easily proved as a direct consequence of well-known characterizations and properties of Sturmian morphisms. We give here a proof based on RC factorizations:

Proposition 2.7. *Let* $\mathbf{w}, \mathbf{w}'$ *be distinct Sturmian words that are fixed points of a nontrivial morphism* $f : A^* \to A^*$*. Then* $\{\mathbf{w}, \mathbf{w}'\} = \{01\mathbf{c}, 10\mathbf{c}\}$*, where* $\mathbf{c}$ *is the characteristic word having the same language.*

Proof. It is easy to see that $\mathbf{w}$ and $\mathbf{w}'$ have the same language. Without loss of generality, we can assume that $\mathbf{w}$ starts with 0 and $\mathbf{w}'$ starts with 1. Let us first show that the combination $\mathbf{w} = 0\mathbf{c}$, $\mathbf{w}' = 1\mathbf{c}$ is not possible. If $f(0\mathbf{c}) = 0\mathbf{c}$ and $f(1\mathbf{c}) = 1\mathbf{c}$, it would follow $\mathbf{c} = \lambda f(\mathbf{c})$ for some word λ, and then $|f(0)| = |f(1)|$, since $\mathbf{c}$ is not ultimately periodic. This is impossible in view of Lemma 2.6.

Thus by Proposition 2.1, $\mathbf{w}$ and $\mathbf{w}'$ are Abelian comparable; let $\mathbf{w} = x_1x_2\cdots$ and $\mathbf{w}' = \widetilde{x_1}\widetilde{x_2}\cdots$ be their corresponding RC factorizations. By contradiction, suppose they contain only words of length > 1. By Proposition 2.4, for all $n \geq 1$ we can write $x_n = 0u_n1$ for suitable words u_n. Since $f(x_1)$ is Abelian equivalent to $f(\widetilde{x_1})$, there must be an $m > 1$ such that $f(x_1) = x_1x_2\cdots x_m$. Hence x_1 and its image both end with 1, so that the word $f(1)$, which clearly starts with 1, ends in 1 as well. As $f(\widetilde{x_1}) = \widetilde{x_1} \cdots \widetilde{x_m}$, by the same argument it follows that $f(0)$ begins and ends in 0. Again, this contradicts Lemma 2.6.

Therefore, there must be some term x_i of length 1 in the RC factorization, that is, (1) holds. As discussed above, this implies that $\mathbf{w}$ and $\mathbf{w}'$ are proximal. Since $\{\mathbf{w}, \mathbf{w}'\} \neq \{0\mathbf{c}, 1\mathbf{c}\}$, by Lemma 2.5 it follows $\{\mathbf{w}, \mathbf{w}'\} = \{\tilde{p}01\mathbf{c}, \tilde{p}10\mathbf{c}\}$ for some $p \in \mathrm{Pref}(\mathbf{c})$; as $\mathbf{w}$ starts with 0 and $\mathbf{w}'$ with 1, the assertion is proved. ☐

Example 2.8. The words $01\mathbf{f}$ and $10\mathbf{f}$ are both fixed by $\tilde{\varphi}^2$, as defined in (4).

3. Main results

3.1. Terms of RC factorizations

We recall the following well-known result by Slater [11], deeply related to the *three distance theorem* proved by Sós [12] (see also [13]):

Theorem 3.1 (Three gap theorem)**.** *Let α be an irrational number with $0 < \alpha < 1$, and let $0 < \beta < 1/2$. The gaps between the successive integers n such that $\{n\alpha\} < \beta$ take either two or three values, one being the sum of the other two.*

Our main theorem shows that RC factorizations have at most 3 distinct terms.

Theorem 3.2. *Let $\mathbf{w}$ be a Sturmian word, and $\mathbf{w} = x_1 x_2 \cdots x_n \cdots$ be an RC factorization of $\mathbf{w}$. The cardinality of the set $X = \{x_n \mid n > 0\}$ is either 2 or 3, and in the latter case, the longest element of X is obtained by concatenating the other two.*

Proof. Since Sturmian words are not periodic, the set X has cardinality at least two. Let $\mathbf{w}' = \widetilde{x_1}\widetilde{x_2} \cdots$, and suppose first that $\mathbf{w}$ and $\mathbf{w}'$ are both lower mechanical words, so that $\mathbf{w} = \mathbf{s}_{\alpha,\rho}$ and $\mathbf{w}' = \mathbf{s}_{\alpha,\rho'}$ for some $\rho, \rho' \in [0, 1[$. Without loss of generality, we may suppose $\alpha < 1/2$ (otherwise it suffices to exchange the roles of the letters 0 and 1), and $\beta := \{\rho' - \rho\} \leq 1/2$ (otherwise we swap $\mathbf{w}$ and $\mathbf{w}'$). Hence $\alpha < 1 - \alpha$ and $\beta \leq 1 - \beta$.

We distinguish two possibilities:

Case 1. If $\alpha \leq \beta$, then $\mathbf{w}$ and $\mathbf{w}'$ cannot be 1 at the same time, i.e., there is no n for which $\{n\alpha + \rho\}$ and $\{n\alpha + \rho'\}$ are both larger than $1 - \alpha$. Assuming, in view of Proposition 2.4, that all terms x_k ($k \geq 1$) are lower Christoffel words (the "upper" case being similar), this implies $X \subseteq \{0\} \cup \{0^k 1 \mid k > 0\}$, since any other lower Christoffel word x would have a 1 in the same position as in $\tilde{x}$. Let i, j, k_1, and k_2 be positive integers such that $i < j$, $x_i = 0^{k_1}1$, and $x_j = 0^{k_2}1$. This implies, since $\mathbf{w}$ and $\mathbf{w}'$ have the same language, that $\mathbf{w}$ has the two factors

$$\begin{aligned} x_i x_{i+1} \cdots x_{j-1} x_j &= 0^{k_1} 1 x_{i+1} \cdots x_{j-1} 0^{k_2} 1\,, \text{ and} \\ \widetilde{x_i}\widetilde{x_{i+1}} \cdots \widetilde{x_{j-1}}\widetilde{x_j} &= 1 0^{k_1} \widetilde{x_{i+1}} \cdots \widetilde{x_{j-1}} 1 0^{k_2} \end{aligned}$$

so that $|k_1 - k_2| \leq 1$ as a consequence of the balance property. By the arbitrary choice of i and j, it follows $X \subseteq \{0, 0^h 1, 0^{h+1} 1\}$ for some $h > 0$, which settles this case.

Case 2. If $\alpha > \beta$, $\mathbf{w}$ and $\mathbf{w}'$ differ exactly in all positions n such that $\{n\alpha + \rho\} \in I_1 \cup I_2$, with $I_1 = [1 - \alpha - \beta, 1 - \alpha[$ and $I_2 = [1 - \beta, 1[$. Note that if $\{n\alpha + \rho\} \in I_1$, then $\{(n+1)\alpha + \rho\} \in I_2$. Hence, if $\rho \notin I_2$ we derive $X \subseteq \{0, 1, 01\}$.

If $\rho \in I_2$, let $(n_i)_{i\geq 0}$ be the increasing sequence of all positive integers such that $\{n_i\alpha+\rho\} \in I_2$. For all $i \geq 0$ we have $\{n_i\alpha + \rho + \beta\} < \beta$, so that by Theorem 3.1 it follows that the set $\{n_{i+1} - n_i \mid i \geq 0\}$ is contained in $\{k_1, k_2, k_1 + k_2\}$ for some distinct integers k_1, k_2 (both greater than 1, as $\alpha < 1 - \beta$). For all $i \geq 0$, x_i is the factor of $\mathbf{w}$ starting with the n_i-th letter and ending in the $(n_{i+1} - 1)$-th one, since it can be written as $1u0$ for some u such that the corresponding factor of $\mathbf{w}'$ is $0u1$. Thus, the elements of X may have length k_1, k_2, or $k_1 + k_2$. There cannot be two of the same length, since any factor of $\mathbf{w}$ which is an upper Christoffel word of length ≥ 2 can be written as $1u0$ where u is a palindromic prefix of the characteristic word of slope α. Hence X has cardinality 2 or 3.

Let now $X = \{y_1, y_2, z\}$ with $|z| = k_1 + k_2$ and $|y_i| = k_i$ for $i = 1, 2$; we can write $z = 1v0$ and $y_i = 1u_i0$. Since u_1, u_2, and v are all palindromic prefixes of the characteristic word, u_i is a prefix and a suffix of v for $i = 1, 2$. Hence

$$v = u_1abu_2 = u_2bau_1 \tag{5}$$

for some letters $a, b \in \{0, 1\}$. If $a \neq b$, it follows either $z = y_1y_2$ or $z = y_2y_1$ and we are done.

By contradiction, let us then suppose $a = b$. From $\alpha < 1/2$ we derive that $a = b = 0$, and that both u_1 and u_2 start with the letter 0. If both were a power of 0 we would reach a contradiction, as no point of the (dense) sequence $\{n\alpha + \rho\}$ would lie in the nonempty interval $[1 - \alpha, 1 - \beta[$. Hence there exists $j \in \{1, 2\}$ such that u_j contains both 0 and 1, so that by Corollary 1.2 either u_j01 or u_j10 is a prefix of v. This is a contradiction, since we are assuming (5) with $a = b = 0$.

When $\mathbf{w}$ and $\mathbf{w}'$ are both upper mechanical words, the proof is symmetrical.

Let us now suppose that both $\mathbf{w}$ and $\mathbf{w}'$ are singular, and that only one of them is a lower mechanical word. By contradiction, suppose that X has more than 3 elements. This means that there exists $j > 3$ such that the set $\{x_1, x_2, \ldots, x_j\} \subseteq X$ has cardinality 4. The word $r = x_1 \cdots x_j$ is prefix of infinitely many nonsingular Sturmian words of slope α, and so is $r' := \widetilde{x_1} \cdots \widetilde{x_j}$. Let $\mathbf{t}, \mathbf{t}'$ be two such nonsingular extensions, of r and r' respectively. The RC factorization of $\mathbf{t}$ and $\mathbf{t}'$ has more than 3 distinct terms, but $\mathbf{t}$ and $\mathbf{t}'$ are both lower mechanical words, a contradiction because of what we proved above. The same argument by contradiction proves that if X has cardinality 3, then its longest element has to be a concatenation of the other two. ☐

3.2. RC factorizations as Sturmian or 3-iet words

Theorem 3.2 allows to consider RC factorizations as infinite words on the *finite* alphabet X. In this section we analyze the structure of such words.

We recall that two finite words u, v are *conjugate* if $u = \lambda\mu$ and $v = \mu\lambda$ for some words λ, μ. The following characterization of Sturmian morphisms was proved in [14, Theorem A.1]:

Theorem 3.3 (See Berthé et al. [14])**.** *A morphism $f : A^* \to A^*$ is Sturmian if and only if it maps the three Christoffel words* 01, 001, *and* 011 *to conjugates of Christoffel words.*

Corollary 3.4. *If u, v, and uv are Christoffel words, then the morphisms $0 \mapsto u, 1 \mapsto v$ and $0 \mapsto v, 1 \mapsto u$ are Sturmian.*

Proof. Immediate consequence of Proposition 1.3 and the previous theorem. ☐

Let us recall a further result ([6, Proposition 2.3.2]) on the Sturmian morphisms φ and $\tilde{\varphi}$ from (4).

Proposition 3.5 (See Berstel and Séébold [6]). *Let* $\mathbf{w}$ *be an infinite word.*

1. *If* $\varphi(\mathbf{w})$ *is Sturmian, then so is* $\mathbf{w}$.
2. *If* $\tilde{\varphi}(\mathbf{w})$ *is Sturmian and* $\mathbf{w}$ *starts with* 0, *then* $\mathbf{w}$ *is Sturmian.*

Given a set $X \subseteq A^*$ and a word $w \in X^*$, by an abuse of language we say that a factorization $w = x_1x_2\cdots$, with $x_i \in X$ for all i, is a *word over the alphabet* X, identifying it with the word $x_1x_2\cdots$ in the free monoid over X (or with the word $f^{-1}(x_1)f^{-1}(x_2)\cdots$, where f is a bijection from a new alphabet B to X). The same identification is made also for factorizations of infinite words.

A *complete return* to $v \in A^*$ is a finite word containing exactly two occurrences of v, one as a prefix and one as a suffix. If w is a finite or infinite word and v is a factor of w, then a (right) *return word to* v *in* w is a word r such that $rv \in \mathrm{Fact}(w)$ is a complete return to v. Left returns can be defined similarly, i.e., replacing rv with vr in the definition.

The following result is a known consequence of a theorem by Vuillon [15] characterizing Sturmian words as the ones having exactly two return words for each factor:

Proposition 3.6 (See e.g. [16]). *Let* p *be a prefix of a Sturmian word* $\mathbf{w}$. *The sequence of return words to* p *in* $\mathbf{w}$ *is Sturmian; that is:*

1. p *has exactly two distinct return words in* $\mathbf{w}$, *say* u *and* v, *and*
2. *given the morphism* $f_p : 0 \mapsto u, 1 \mapsto v$, *the word* $\mathbf{w}_p \in A^\omega$ *such that* $f_p(\mathbf{w}_p) = \mathbf{w}$ *is Sturmian.*

We can now begin to shed light on the structure of RC factorizations.

Proposition 3.7. *Let* $\mathbf{w} = x_1x_2\cdots$ *and* X *be defined as in Theorem 3.2. Suppose* $X = \{u, v, z\}$ *with* $z = uv$. *Then:*

1. *The factorization* $\mathbf{w} = y_1y_2\cdots$, *obtained from the starting RC factorization by replacing each occurrence of* z *with* $u \cdot v$ *(that is, defined so that for all* i *with* $x_i = z$, *there exists* j *with* $y_1\cdots y_{j-1} = x_1\cdots x_{i-1}$, $y_j = u$, *and* $y_{j+1} = v$) *is also reversible Christoffel.*
2. *The new factorization* $y_1y_2\cdots$ *is also a Sturmian word on the alphabet* $\{u, v\}$.

Proof. The result is trivially verified when $\{u, v\} = \{0, 1\}$. Let us then suppose this is not the case; by Proposition 1.3 we deduce that either u is a prefix of v, or v is a suffix of u. Distinguishing such two cases, we shall first prove our second claim, i.e., that the new factorization defines a Sturmian word $\hat{\mathbf{w}} = f^{-1}(y_1)f^{-1}(y_2)\cdots$, where $f : 0 \mapsto u, 1 \mapsto v$.

- If u is a prefix of v, let $n > 0$ be the greatest integer such that u^n is a prefix of v. Clearly u^n is a prefix of $\mathbf{w}$; we shall prove that u and v are the return words to u^n in $\mathbf{w}$, thus showing that $\hat{\mathbf{w}}$ is indeed Sturmian by Proposition 3.6.

  It is easy to check that u^{n+1} and vu^n are indeed factors of $\mathbf{w}$. Since u is primitive, it is clearly a return word to u^n. Let $v = u^nu'$ for some $u' \in A^*$. By Proposition 1.3, u' and uu' are Christoffel words, so that we have either $\{u, u'\} = \{a, b\}$ or $u' \in \mathrm{Suff}(u)$ (u cannot be a prefix of u' by the maximality of n). Clearly u^n is a prefix and a suffix of $vu^n = u^nu'u^n$; we need to show that has no other occurrences. If $\{u, u'\} = A$, this is trivial. If u' is a suffix of u, then u^n cannot have internal occurrences in $u^nu'u^n$ because u is unbordered. Therefore u and v are return words to u^n in $\mathbf{w}$, so that $\hat{\mathbf{w}}$ is Sturmian.

- If v is a suffix of u, let $m > 0$ be the greatest integer such that v^m is a suffix of u. The same argument as above shows that $v^m u$ and v^{m+1} are complete returns to v^m. Thus, writing $u = v'v^m$ for some $v' \in A^*$, we get that v and $v^m v'$ are the return words to v^m in the word $v^m\mathbf{w}$. The sequence of these return words is again determined by $\hat{\mathbf{w}}$, as $v^m\mathbf{w} = g(\hat{\mathbf{w}})$ with $g : 0 \mapsto v^m v', 1 \mapsto v$. Hence, to prove that $\hat{\mathbf{w}}$ is Sturmian, by Proposition 3.6 we only need to show that $v^m\mathbf{w}$ is Sturmian.

  By mirroring the argument used in the proof of Proposition 2.1, we get that since $\{\mathbf{w}, \mathbf{w}'\} \neq \{0\mathbf{c}, 1\mathbf{c}\}$, there exist arbitrarily long Abelian equivalent words r, r' such that $r\mathbf{w}$ and $r'\mathbf{w}'$ are Sturmian. By Theorem 3.2, the set of terms in the RC factorization of $r\mathbf{w}$ determined by $r'\mathbf{w}'$ cannot be larger than $X = \{u, v, z\}$. Since $u = v'v^m$ and $z = v'v^{m+1}$, any sufficiently long r will have v^m as a suffix. We have thus proved that $v^m\mathbf{w} = g(\hat{\mathbf{w}})$ is Sturmian, and so is $\hat{\mathbf{w}}$.

Since $\hat{\mathbf{w}}$ is Sturmian, and the morphism $\tilde{f} : 0 \mapsto \tilde{u}, b \mapsto \tilde{v}$ is Sturmian by Corollary 3.4, the word $\tilde{f}(\hat{\mathbf{w}})$ is Sturmian too, so that the new factorization is actually the RC factorization of $\mathbf{w}$ determined by $\tilde{f}(\hat{\mathbf{w}})$. This completes the proof. □

Remark 3.8. If u, v, and uv are the terms of an RC factorization $\mathbf{w} = x_1x_2\cdots$, then by Proposition 1.3 exactly one among u^2v and uv^2 is a factor of $\mathbf{w}$. As a consequence of Proposition 3.7, it is easy to see that if u^2v (resp. uv^2) is a factor of $\mathbf{w}$, then each occurrence of v (resp. u) in the factorization is preceded by u (resp. followed by v), provided that $x_1 \neq v$. This gives rise to the following "converse" of Proposition 3.7.

Proposition 3.9. *Let u, v, and $\mathbf{w} = x_1x_2\cdots$ be as above. Replacing in the factorization each occurrence of $u \cdot v$ with one of $z = uv$ (that is, defining y_n for $n > 0$ so that for all i where $x_i = u$ and $x_{i+1} = v$ there exists j with $x_1\cdots x_{i-1} = y_1\cdots y_{j-1}$ and $y_j = z$) produces a new RC factorization of $\mathbf{w}$, which is also a Sturmian word (over $\{z, u\}$ or $\{z, v\}$).*

Proof. Let us first assume $u^2v \in \mathrm{Fact}(\mathbf{w})$. By the previous Remark, it is clear that the replacement yields a factorization of $\mathbf{w}$ where v does not appear. Hence, the new factorization can be seen as an infinite word on the alphabet $\{u, z\}$, i.e., the image of some word $\bar{\mathbf{w}} \in A^\omega$ under the morphism $h : 0 \mapsto z, 1 \mapsto u$. To see that $\bar{\mathbf{w}}$ is Sturmian, by Proposition 3.5 it suffices to observe that $\varphi(\bar{\mathbf{w}}) = \hat{\mathbf{w}}$, where $\hat{\mathbf{w}}$ is given by Proposition 3.7 and φ is the Fibonacci morphism as in (4).

To show that this new factorization is actually RC, once again we just need to observe that it is obtained by Abelian comparison with the word $\tilde{h}(\bar{\mathbf{w}})$, where the morphism $\tilde{h} : 0 \mapsto \tilde{z}, 1 \mapsto \tilde{u}$ is Sturmian by Corollary 3.4.

Now suppose $uv^2 \in \mathrm{Fact}(\mathbf{w})$ instead. Essentially the same argument as above applies; we consider the new factorization as the image of an infinite word $\bar{\mathbf{w}}$ under the morphism $h : 0 \mapsto z, 1 \mapsto v$. We have $(E \circ \tilde{\varphi})(\bar{\mathbf{w}}) = \hat{\mathbf{w}}$, and $\bar{\mathbf{w}}$ starts with 0, so that $\bar{\mathbf{w}}$ is Sturmian by Proposition 3.5; the factorization is RC since it is obtained by Abelian comparison with $\tilde{h}(\bar{\mathbf{w}})$, where $\tilde{h} : 0 \mapsto \tilde{z}, 1 \mapsto \tilde{v}$ is Sturmian by Corollary 3.4. □

We recall (cf. [17]) that a *3-iet word* is an infinite word coding the orbit of a point ρ under a *three-interval exchange transformation* T. More precisely, given an interval $I = [0, \ell] \subseteq \mathbb{R}$ containing ρ and subdivided in three intervals $I_a = [0, \alpha]$, $I_b = [\alpha, \alpha + \beta]$, and $I_c = [\alpha + \beta, \ell]$, we let T be the piecewise linear transformation of I exchanging the three subintervals according to the permutation (321), i.e., let $T : I \to I$ be defined by $T(\xi) = \xi + t_x$ if $\xi \in I_x$, where $x \in \{a, b, c\}$ and

$$t_a = \ell - \alpha, \quad t_b = \ell - 2\alpha - \beta, \quad \text{and } t_c = -\alpha - \beta\ .$$

The 3-iet word determined by α, β, ℓ, and ρ is then the infinite word indexed over $\mathbb{N}$ whose n-th letter is $x \in \{a, b, c\}$ if $T^n(\rho) \in I_x$.

Let $\sigma, \sigma' : \{a, b, c\}^* \to A^*$ be morphisms defined by

$$\sigma(a) = 0 = \sigma'(a), \quad \sigma(b) = 01, \ \sigma'(b) = 10, \quad \sigma(c) = 1 = \sigma'(c) \, . \tag{6}$$

The following result was proved in [17, Theorem A]:

Theorem 3.10 (See Arnoux et al. [17]). *An infinite word* $\mathbf{u}$ *on the alphabet* $\{a, b, c\}$*, whose letters have positive frequencies, is an aperiodic 3-iet word if and only if* $\sigma(\mathbf{u})$ *and* $\sigma'(\mathbf{u})$ *are Sturmian words.*

As a consequence, we get that RC factorizations are in general 3-iet words:

Corollary 3.11. *Let* $\mathbf{w}$ *be a Sturmian word, and* $\mathbf{w} = x_1 x_2 \cdots$ *be an RC factorization with* $X = \{x_n \mid n > 0\} = \{u, v, z\}$, $z = uv$*. If every word of* X *occurs more than once in the factorization, then* $x_1 x_2 \cdots$ *is an aperiodic 3-iet word over the alphabet* $\{u, z, v\}$*.*

Proof. Let $\tau : \{a, b, c\} \to \{u, v, z\}$ be defined by $\tau(a) = u$, $\tau(b) = z$, and $\tau(c) = v$. We need to show that the infinite word $\dot{\mathbf{w}} := \tau^{-1}(x_1)\tau^{-1}(x_2) \cdots$ is a 3-iet word. Clearly, exchanging the roles of $\mathbf{w}$ and $\mathbf{w}'$ (and letting $\tau(a) = \tilde{u}$ etc.) does not change $\dot{\mathbf{w}}$; therefore, without loss of generality, in the following we can assume by Proposition 2.4 that X is made of lower Christoffel words.

Let α be the slope of $\mathbf{w}$ and $\mathbf{w}'$, and let ρ and ρ' respectively be their intercepts. As is well-known, any factor γ of $\mathbf{w}$ corresponds to an interval I_γ on the unit circle, i.e., γ occurs at position n in $\mathbf{w}$ if and only if $\{n\alpha + \rho\} \in I_\gamma$; moreover, since $\mathbf{w}$ and $\mathbf{w}'$ have the same slope, the positions of $\tilde{\gamma}$ in $\mathbf{w}'$ are identified by

$$\{n\alpha + \rho'\} \in I_{\tilde{\gamma}} \iff \{n\alpha + \rho\} \in I_{\tilde{\gamma}} - \rho' + \rho$$

where $I_{\tilde{\gamma}} - \rho' + \rho$ is a translation on the unit circle (i.e., the sum is taken modulo 1).

Let now $\gamma \in X$, and suppose first that $|\gamma| > 1$. The relation

$$\{n\alpha + \rho\} \in I_\gamma \cap (I_{\tilde{\gamma}} - \rho' + \rho) \tag{7}$$

identifies the positions n of all occurrences of γ in $\mathbf{w}$ such that $\tilde{\gamma}$ occurs at the same position in $\mathbf{w}'$. As $|\gamma| > 1$, we have $\gamma = 0q1$ and $\tilde{\gamma} = 1q0$ for some word q. We claim that all occurrences of γ whose position satisfies (7) appear in the RC factorization. Indeed, if one of such occurrences did not correspond to x_i for any $i \geq 1$, then the first 0 of $0q1$ (resp. 1 of $1q0$) would have to be the last letter of some x_j (resp. $\widetilde{x_j}$), against the fact that all x_j's are lower Christoffel words.

Now suppose γ is a letter. Without loss of generality, we can assume $\alpha < 1/2$. If $\gamma = 1$, then necessarily $X = \{0, 1, 01\}$, so that (7) again identifies exactly all occurrences of γ in $\mathbf{w}$ that appear in the RC factorization. Let then $\gamma = 0$, so that X is $\{0, 0^n1, 0^{n+1}1\}$ for some $n \geq 0$. As a consequence of the balance condition, it is easy to see that a position n where 0 occurs in both $\mathbf{w}$ and $\mathbf{w}'$ corresponds to x_i for some $i \geq 1$ if and only if it is followed in $\mathbf{w}$ by $x_{i+1} = 0^n1$. Hence, such positions are exactly those that satisfy

$$\{n\alpha + \rho\} \in I_{0^{n+1}1} \cap (I_{010^n} - \rho' + \rho) \, . \tag{8}$$

In all cases and for all $\gamma \in X$, we have identified intervals corresponding to the occurrences of γ in the RC factorization, namely the ones in (7) or (8). By hypothesis, γ occurs at least twice in the factorization; by the irrationality of α, such intervals must then have nonempty interior, so that the gaps between consecutive integers n satisfying (7) or (8) are bounded. Thus, every term in the RC factorization, and so every letter in $\dot{\mathbf{w}}$, occurs with positive frequency.

By Theorem 3.10, it remains to prove that the two words $\sigma(\dot{\mathbf{w}})$, $\sigma'(\dot{\mathbf{w}})$ are Sturmian, where σ and σ' are the morphisms defined in (6). In fact, it is easy to check that $\sigma(\dot{\mathbf{w}}) = \hat{\mathbf{w}}$ and $\sigma'(\dot{\mathbf{w}}) = \hat{\mathbf{w}}'$, i.e., the words $\sigma(\dot{\mathbf{w}})$ and $\sigma'(\dot{\mathbf{w}})$ coincide with the Sturmian words obtained in the proof of Proposition 3.7, applied respectively to the RC factorizations $\mathbf{w} = x_1 x_2 \cdots$ and $\mathbf{w}' = \widetilde{x_1}\widetilde{x_2}\cdots$. The result follows. □

Remark 3.12. The hypothesis that every word of X occurs more than once in the factorization is necessary. For instance, it is easy to see that for the RC factorization $x_1 x_2 \cdots x_n \cdots$ of $\mathbf{w} = 010\mathbf{f}$ determined by $\mathbf{w}' = 1001\mathbf{f}$, one has $X = \{0, 01, 001\}$, but the term 0 occurs exactly once, as x_2. In fact, by Proposition 3.7, the word $\mathbf{f} = x_3 x_4 \cdots$ is Sturmian over the alphabet $\{01, 001\}$.

4. Future work

We believe that much can still be said about reversible Christoffel factorizations; for example, it would be interesting to characterize the set of terms X in terms of the slope α and the difference between the intercepts, in particular to distinguish when the cardinality of X is 2 or 3.

Acknowledgment

The second author would like to thank the Department of Mathematics at the University of Turku for its support during his part-time employment there in 2011–2012 as a member of the FiDiPro unit.

References

[1] M. Barge, B. Diamond, Coincidence for substitutions of Pisot type, Bull. Soc. Math. France 130 (2002) 619–626.
[2] M. Bucci, A. De Luca, L. Q. Zamboni, Reversible Christoffel factorizations, in: First Russian-Finnish Symposium on Discrete Mathematics (RuFiDiM), Euler International Mathematical Institute, St. Petersburg, Russia, 2011, pp. 13–15.
[3] M. Morse, G. A. Hedlund, Symbolic dynamics II. Sturmian trajectories, Amer. J. Math. 62 (1940) 1–42.
[4] J. Berstel, A. de Luca, Sturmian words, Lyndon words and trees, Theoret. Comput. Sci. 178 (1997) 171–203.
[5] A. de Luca, Sturmian words: structure, combinatorics, and their arithmetics, Theoret. Comput. Sci. 183 (1997) 45–82.
[6] J. Berstel, P. Séébold, Sturmian words, in: M. Lothaire (Ed.), Algebraic Combinatorics on Words, Cambridge University Press, Cambridge UK, 2002. Chapter 2.
[7] J. Berstel, A. Lauve, C. Reutenauer, F. Saliola, Combinatorics on Words: Christoffel Words and Repetition in Words, volume 27 of *CRM monograph series*, American Mathematical Society, 2008.
[8] M. Bucci, S. Puzynina, L. Q. Zamboni, Central sets defined by words of low factor complexity, 2011. ArXiv:1110.4225.
[9] S. Yasutomi, On Sturmian sequences which are invariant under some substitutions, in: Number theory and its applications (Kyoto, 1997), volume 2 of *Dev. Math.*, Kluwer Acad. Publ., Dordrecht, 1999, pp. 347–373.
[10] V. Berthé, H. Ei, S. Ito, H. Rao, On substitution invariant Sturmian words: an application of Rauzy fractals, Theor. Inform. Appl. 41 (2007) 329–349.
[11] N. Slater, The distribution of the integers n for which $\{\theta n\} < \phi$, Proc. Cambridge Philos. Soc. 46 (1950) 525–534.
[12] V. Sós, On the distribution mod 1 of the sequence $n\alpha$, Ann. Univ. Sci. Budapest, Eötvös Sect. Math 1 (1958) 127–134.
[13] P. Alessandri, V. Berthé, Three distance theorems and combinatorics on words, Enseign. Math. (2) 44 (1998) 103–132.
[14] V. Berthé, A. de Luca, C. Reutenauer, On an involution of Christoffel words and Sturmian morphisms, European J. Combin. 29 (2008) 535–553.

[15] L. Vuillon, A characterization of Sturmian words by return words, European J. Combin. 22 (2001) 263–275.
[16] J. Justin, L. Vuillon, Return words in Sturmian and episturmian words, Theor. Inform. Appl. 34 (2000) 343–356.
[17] P. Arnoux, V. Berthé, Z. Masáková, E. Pelantová, Sturm numbers and substitution invariance of 3iet words, Integers 8 (2008) A14, 17.